IEEE Open Journal of the Computer Society

Received 1 April 2024; accepted 8 April 2024. Date of publication 18 April 2024;
date of current version 2 May 2024. The review of this article was arranged by Associate Editor Song Guo.

Digital Object Identifier 10.1109/OJCS.2024.3389781

Optimal Neighborhood Contexts in Explainable AI: An Explanandum-Based Evaluation

URJA PAWAR [1], DONNA O'SHEA [1] (Senior Member, IEEE), RUAIRI O'REILLY [1] (Senior Member, IEEE), MAEBH COSTELLO [2], AND CHRISTIAN BEDER [1]

[1]Munster Technological University (MTU), T12P928 Cork, Ireland
[2]Mckesson, T12XN72 Cork, Ireland

CORRESPONDING AUTHOR: URJA PAWAR (e-mail: urja.pawar@mycit.ie).

This work was supported by Science Foundation Ireland and Mckesson Pvt. Ltd.
This article has supplementary downloadable material available at https://doi.org/10.1109/OJCS.2024.3389781, provided by the authors.

ABSTRACT Over the years, several frameworks have been proposed in the domain of Explainable AI (XAI), however their practical applicability and utility need to be clarified. The neighbourhood contexts are shown to significantly impact the explanations generated by XAI frameworks, thus directly affecting their utility in addressing specific questions, or "explananda". This work introduces a methodology that use a comprehensive range of neighbourhood contexts to evaluate and enhance the utility of specific XAI techniques, particularly Feature Importance and CounterFactuals. In this evaluation, two explananda are targeted. The first one examines whether features' collection should be halted as per the AI model based on the sufficiency of the current set of information. Here, the information refers to the features present in the data used to train the AI-based system. The second one explores what is the most effective information (features) that should be collected next to ensure that the AI outputs the same classification as it would have generated with all the information present. These questions serve as a platform to demonstrate our methodology's ability to assess the impact of customised neighbourhood contexts on the utility of XAI.

INDEX TERMS Explainable AI, healthcare, SHAP, LIME, counterfactuals, DiCE ML, evaluation.

I. INTRODUCTION

In recent years, efforts to demystify the Artificial Intelligence(AI) and Machine Learning(ML) models has led to the development of Explainable AI(XAI) techniques, such as Feature Importance(FI) and CounterFactuals(CFs). In XAI, an "explanation" addresses a specific question called explanandum (plural - explananda) about ML models to enhance our understanding of them [1]. However, the interpretations of the provided explanations remain unclear, leading to their potential misapplication or over-extension to other explananda. There is a pressing need to critically evaluate the precise applicability of explanations to prevent misinterpretations and harness their potential effectively [2].

To assess the applicability and reliability of XAI framework, three distinct evaluation strategies are followed - application-grounded, human-grounded and functionally-grounded, as proposed by [3]. Application-grounded evaluations have been conducted by researchers [4], [5] to evaluate the practical applicability of explanations in a given domain using domain-relevant explananda. Human-grounded evaluations have been conducted through user studies in [6], [7] to assess explananda related to user comprehension. Functionally-grounded evaluations have also been conducted in [8], [9] to assess the explanations on the technical explananda such as robustness, and faithfulness.

However, the explanations generated by XAI technique are sensitive to the sample distributions in a local neighborhood that is used for the analysis [10], [11]. These distributions are referred to as "contexts" in this paper. If an XAI framework addresses a specific explanandum, changing the neighborhood context will impact the efficacy of that framework for the given explanandum [12]. While various evaluation approaches

have been proposed in the literature, a gap exists in comprehensively evaluating the impact of various neighborhood contexts on the effectiveness of explanations to address a given explanandum.

In this work, we propose a methodology that provided a structured approach of facilitating explanandum-based evaluations of XAI using neighbourhood contexts. We have selected two application-grounded explananda as use cases to demonstrate our methodology. These explananda, while applicable across multiple domains, are particularly valuable in enhancing ML-based Clinical Decision Support Systems (CDSS) based on our review of the literature [13], [14].

To identify suitable explananda, we focused on the challenges faced in medical diagnostics by CDSS, especially when relying on tabular data. Typically, these systems assume access to a complete set of features, but this is often not feasible in low-resource settings [15]. Based on this challenge, the first explanandum, termed "early stopping," is - whether a medical workflow can be halted early based on XAI explanations. This is useful when clinicians are confident enough to make a diagnosis without collecting more information about a patient and want to assess whether the AI system aligns with this thought process [14]. The second, "next best feature," is - what should be the next critical medical information or a feature needed for an accurate diagnosis by the AI system. Applying standard XAI techniques may not provide clear answers for these explananda. For example, highly-ranked features by FI do not necessarily suggest that they are the only ones needed for a diagnosis by the AI system [16].

The importance of this work is two-fold: different neighbourhoods are explored to understand their impact on the utility of the explanations for specific explananda. Additionally, it provides an enhanced interpretation of XAI explanations generated using different contexts and brings awareness to their limitations. The methodology is applicable for tabular datasets as they have distinct features that can be collected using a workflow. For evaluation, we considered state-of-the-art XAI techniques - SHapley Additive exPlanations (SHAP) and Locally Interpretable Model-agnostic Explanations (LIME), and DiCE(Diverse Counterfactual Explanations); that are not limited to explaining gradient-based models and are applicable to all types of ML models. For "early stopping", our evaluation considered SHAP and LIME because these frameworks can highlight features to indicate whether they are sufficient for a reliable classification [16]. CFs do not inherently focus on feature sufficiency and are not included in the evaluation for this explanandum. However, all three techniques – SHAP, LIME, and DiCE – are assessed under the explanandum of "next best feature" as all of them highlight features that can influence classification. The key contributions of our work are outlined.

- Formalized definitions of the explananda relevant to the medical domain are proposed.
- Experimental methodology to evaluate XAI frameworks with respect to explananda is proposed
- Deriving optimal neighbourhood contexts to achieve better performance of XAI on the explananda is demonstrated.

II. BACKGROUND INFORMATION

In this section, aligned to the core contribution of the work, background information on contexts is provided. As we evaluated neighborhood contexts for individual XAI frameworks, we also discuss the mechanism of XAI frameworks and detail how an input neighborhood context is used by the frameworks to generate explanations.

A. CONTEXTS

In XAI, explanations are usually categorised as either local or global [17]. Global explanations provide a broad understanding of the ML model, while local explanations focus on specific model decisions for individual inputs. In this work, the selected explananda focus on patient-specific medical workflows; therefore, local explanations are considered for the evaluation. Local explanations often use neighbourhoods (here, a set of medical records) around an input sample (a medical record) to generate the explanation [10]. This neighbourhood can contain training data samples as well as perturbed samples by changing feature values in the input sample using specific criteria.

In XAI, a neighbourhood refers to samples close to an input, while a "context" c specifies the subset of these samples selected for analysis based on a criteria [16]. A context is distinct from the local neighbourhood as it is not just a collection of locally similar samples, but rather those selected using a specific criteria, forming a distribution to be used for generating explanations [18]. This distribution can be characterised in different ways, such as marginal or conditional distributions, indicative of how feature values have been perturbed or intervened upon. A *standard* context is defined as the distribution of the training data. The mathematical formulations of contexts are not the main focus of this paper, we assume that these contexts are appropriately defined based on their respective criteria.

Fig. 1 shows two example contexts using dotted blue curves. The selected samples in the contexts are specifically colored to distinguish their types: training samples by yellow, and the perturbed ones by grey. In context $c1$, the classification of an input sample (highlighted in red) in the green category is analysed by using both training (yellow) and perturbed samples (grey). $c1$'s criteria is to select neighbourhood samples that share the same classification as the input sample. The distribution of $c1$ can be represented by conditional probability based on the likelihood of that the samples are within the same class. Context $c2$ generates explanations by selectively comparing the input to the neighbouring training samples (yellow) from the opposite class (blue). In $c2$, the analysis gives more weight to samples nearer to the input, as shown by their size, with their distribution defined by a weighted conditional probability. Various contexts and their criteria are elaborated in Section IV.

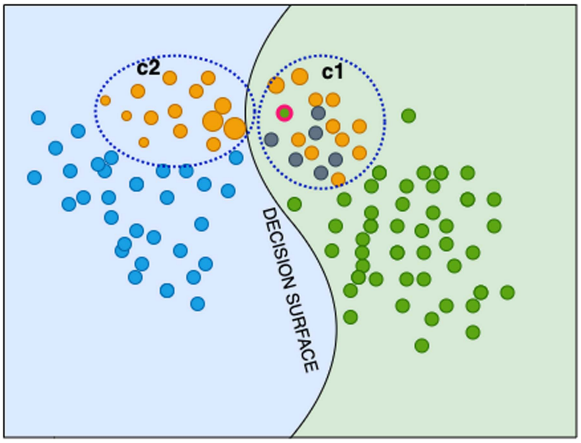

FIGURE 1. Contexts c1 and c2 highlighted by dotted blue curves. Yellow points represent selected training samples and grey points represent the perturbed samples.

B. XAI FRAMEWORKS

FI and CFs are predominant XAI techniques [17]. FI assigns scores to features called FI scores with positive or negative values to denote whether a feature positively or negatively impacts a classification [19]. FI reflects the idea of "sufficiency" to highlight features that are sufficient in maintaining a classification [18]. CFs describe minimal feature changes required to alter a classification [20]. In CFs a feature is deemed highly important if it is frequently changed to alter the outcome. CFs align with the concept of "necessity", where a feature change could alter a classification [16].

For model-agnostic FI, SHAP[1] and LIME[2] are state-of-the-art open-source FI frameworks [21], [22]. Other FI frameworks such as integrated gradients and layer-wise relevance propagation are specific to neural networks and mainly used for image datasets [23].

SHAP: To explain an input-output pair $(u, f(u))$, SHAP [24] estimates the average marginal contribution of an individual j^{th} feature considering all possible feature subsets. The feature subsets are constructed by including the original value of the features belonging to the subset, and replacing the values of other features from the training data samples. Afterwards, the prediction difference by $\hat{f}$ on including and excluding the j^{th} feature value in different feature subsets is analysed. Here, excluding means replacing that feature value from the ones in the training data [25]. A kernel function is used to weigh the samples based on the size of the feature subset used, i.e., how many feature values are common between the original input and the sample. The *standard* context for SHAP for explaining an input is a constructed set of perturbed samples that contain a combination of feature values from the original input and training samples that is inputted to SHAP as a neighborhood context. These perturbed samples with few or too many common feature values are given higher weights based on the kernel function.

LIME: To explain an input-output pair $(u, f(u))$ locally, LIME [10] estimates the FI scores by generating a local neighbourhood around u and training an interpretable linear model g on the neighbourhood samples weighted using a distance metric, and their classifications by f as ground truth. The coefficients of a linear model that best approximates f provide the FI scores. Small changes in feature values with large changes in the model's prediction cause larger FI scores. The *standard* context in LIME uses a local neighbourhood based on the training samples closer to the input as per the Euclidean distance.

CF: Wachter et al. [20] introduced a CF framework that solves an optimization problem to identify instances close to the original input (as per a distance metric) but classified differently. To have more diverse CFs, Mothilal et al. [26] proposed DiCE framework.[3] In this work, we evaluate CFs from DiCE because Wachter framework [20] provide only one CF explanation, while DiCE [26] provides multiple CFs. Multiple CFs can be collectively analysed to generate importance scores for features by counting how many times a feature's value changed in the CFs. The context used by CF frameworks is a local neighborhood usually using the Euclidean distance but restricted to samples having different classification than the input sample. If an additional criterion is added to this context (e.g. not allowing changes in certain features), the selection of a CF explanation will vary. Table 1 summarises the context construction in the three discussed XAI frameworks.

TABLE 1. Summary of Methods for Context Construction

Method	Context Construction
SHAP	Generates perturbed samples by random replacement of feature values using neighborhood samples.
LIME	Uses input neighborhood to train a linear model, with labels from the black-box classifier as ground truth.
DiCE	Uses input neighborhood from opposite side of the decision surface; analyses features' value changes in CFs to calculate their impact on classification.

III. FORMALIZED EXPLANANDA

The sections below detail the explanandum-based evaluation using feature rankings by XAI and their applications. These explananda are used as examples from the medical domain to demonstrate application-based evaluation by simulating scenarios in a domain and evaluating utility of XAI. The explananda are discussed with respect to the medical domain but can be extended to other domains as well.

1) EARLY STOPPING

This evaluation assesses the utility of feature rankings using a given context in enabling an 'early stop' in medical

[1][Online]. Available: https://github.com/slundberg/shap
[2][Online]. Available: https://github.com/marcotcr/lime
[3][Online]. Available: https://github.com/interpretml/DiCE

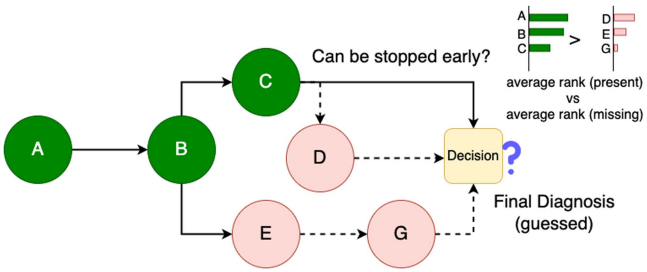

FIGURE 2. Workflow depicted by the arrows will be stopped based on the average rankings of present features - A, B, and C (in green), and missing features - D, E, and G (in red).

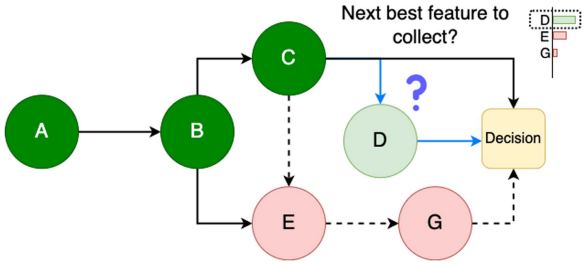

FIGURE 3. The next best feature D is selected to arrive to the final decision based on the importance rankings.

workflows. An 'early stop' signifies that sufficient data has been collected, and additional data is unlikely to change the classification. This decision is crucial in medical settings, where unnecessary tests can cause undue patient stress and increased healthcare costs. The utility of this explanandum could be extended to other domains as well. For example, in autonomous driving, the decision to brake can be taken early if an obstacle is detected using some sensors' inputs, without waiting for further sensors' inputs.

The explanandum is represented using an example as shown in Fig. 2. The evaluation process involves comparing the average feature rankings of collected or "present" features (A, B, C) and the "missing" features (D, E, G). These feature rankings are generated by XAI frameworks for each input sample in which missing features are imputed using mean values. The FI scores generated will reflect the importance of the imputed mean values and will not consider other possible values of the missing features. The proposed methodology aims to use these specific FI scores for the evaluation such that an optimal context can be derived. This optimal context can be specifically used to generate feature rankings in mean-imputed samples such that early stopping can be predicted based on average rank comparison. We used mean imputation for its generic applicability, and straightforwardness [27], [28], [29]. The imputation is done because we aim to assess the direct utility of XAI at a specific stage of a medical workflow with missing features and evaluate the utility of feature ranking in an imputed input sample.

This evaluation examines whether the rankings generated using a given neighborhood context reflect features' sufficiency. In Fig. 2, as shown by the importance bar-plots of present and missing features (in green, and red respectively), clinicians can observe that the already collected features are sufficiently informative to the AI to consider an early stop in the workflow, based on their higher average rank compared to the missing features. For comparison, the average rankings of present and missing features are used to have a holistic view of relative importance, unlike percentile methods that lack depth by only comparing i^{th} percentile features [30]. As detailed in Section IV, the comparison involves the average rank of k missing features and the average rank of top-k present features ensuring that our focus remains on the most influential present features and avoiding dilution of the average rank from lower-ranked ones. In [31], we gathered the workflow information for the heart disease diagnosis based on the medical literature for demonstrating a specific utility of XAI. However, in this work, we used random multiple sets of present/missing features for the evaluation due to the lack of workflow information with respect to other datasets - Cervical Cancer, and Diabetes. Our analysis is generic and further validation with respect to a specific workflow will be considered in future work.

2) NEXT BEST FEATURE

This evaluation assesses the utility of feature rankings in identifying the necessary "next best" feature to collect. If additional data is required for a diagnosis, feature rankings can then be utilised to recommend the specific feature value that would be most informative. This can enable clinicians to validate an ML model's alignment with respect to the medical diagnosis of a patient that will lead to faster, and accurate diagnosis. In this explanandum, that the order of feature collection is arbitrary and is not tied to a specific medical workflow, allowing for broader applicability of this explanandum. For example, in digital marketing, if the initial data about a user (e.g. browsing history) doesn't provide a clear profile, the system can identify the next piece of data (e.g. survey responses, purchase history) to better target ads.

Similar to the early stopping, evaluation is done by generating feature rankings of input samples such that the original values are used for present features, and the mean values for missing features as we don't have their original values at a given stage. The selection of the next best feature is based on the highest ranked missing feature by an XAI framework. This methodology assumes that the feature rankings reflect their information content towards generating a classification and that the features don't need to be collected in a particular sequence.

This explanandum is shown in Fig. 3, where feature D is chosen as the next best feature among missing features D, E, and G. The evaluation uses random combinations of present and missing features.

IV. METHODOLOGY

This section presents our methodology for evaluating XAI frameworks – SHAP, LIME, and DiCE – across different contexts, to identify the optimal context for each framework

TABLE 2. Datasets Information and Accuracy Scores

Dataset Name	Dataset Information		Accuracy	
	no. of records	*features*	SVM	RF
Cervical Cancer [32]	1256	24	97.10	90.88
Heart Disease [33]	297	15	85.18	86.67
Diabetes [34]	768	7	75.65	79.24

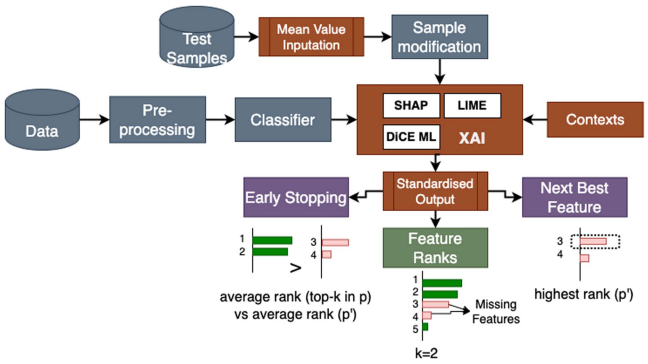

FIGURE 4. Proposed methodology for evaluating XAI.

in each explanandum. We specifically chose three medical datasets – Cervical Cancer, Heart Disease, and Diabetes – primarily because they comprise of individual features that can be gathered sequentially in medical workflows as represented in the explananda's descriptions. For example, a cholesterol test is performed after initial patient information for heart disease diagnosis.

We used a Support Vector Machine (SVM) with a linear kernel for classification. Since our evaluation focuses model-agnostic on XAI frameworks rather than the classifier's performance, we have discussed the results using SVM classifier. To demonstrate the model-agnostic nature of evaluation results, additional results using using the Random Forest (RF) classifier are provided in the supplementary material (Appendix A). Table 2 summarises the dataset details and classifiers' accuracy scores.

The proposed methodology can be used in any domain and on any tabular dataset where features are sequentially collected using workflows. Fig. 4 illustrates our methodology: we select a dataset, preprocess it, and train a classifier. The evaluation begins with selecting input samples and modifying their feature values to simulate scenarios for early stopping and next best feature for XAI utility assessment. Missing features p' are imputed with mean values to mimic data unavailability. A chosen context type is then used to create a local neighbourhood around the modified input sample. The modified sample, along with its neighbourhood and the classifier, are input to the XAI frameworks to produce explanations.

Explanations are then standardised to have consistent evaluation. We used the magnitude of importance scores (FI scores) from XAI to generate feature rankings. While SHAP and LIME provide FI scores directly, in DiCE, the scores are inferred from how often a feature is suggested for change, with more frequently changed features receiving higher FI scores and thus top rankings [16].

Finally, the standardised outputs from XAI frameworks are evaluated against the formalized explananda. In Fig. 4, the standard output is the feature rankings that include red bars showing missing features' FI scores (p') and green for present ones (p). The ranking feeds into the two evaluations: (1) early stopping - where the average rankings of the missing and present features (p', $p(top-k)$) are compared; (2) next best feature - where the highest-ranked missing feature is evaluated.

A. CONTEXTS

As mentioned in Section II, XAI frameworks craft their own neighborhoods from a given input context. We evaluated various contexts with XAI frameworks to test their effectiveness on our chosen explananda. This includes contexts existing in XAI literature [12], [16] as well as newly proposed probabilistic and range-based contexts. The taxonomy of contexts used is described below:

1) *Standard:* This uses samples only from the training set, reflecting training data distribution [24].
2) *Generic:* This includes both - samples from training data and those generated through perturbations [11].
3) *Distance metric based (dist):* This includes samples close to the input, determined using Mahalonobis distance. The Mahalonobis distance is calculated using a covariance matrix based on the training data, ensuring that it reflects the feature relationships and variations seen during the model's training [10].
4) *Restricted Outside (outside):* This has samples with classifications different from the input sample [35].
5) *Restricted Inside (inside):* This has samples that share the same classification as the input sample [36].
6) *Probabilistic (prob):* In this, samples with continuous feature values near the input sample are given higher likelihood based on a linear function of their proximity. For example, if the input sample has an age value of 25, then ages 24 and 26 are more likely to be included than ages 20 or 30. This method allows for the inclusion of samples further from the input but with progressively lesser weight, distinguishing it from strictly distance-based methods. Probabilities for categorical/binary data are uniform. There is no inter-feature relationship considered in this context. This context is proposed to analyse the model's behaviour across various samples that are closely clustered around feature values similar to the input.
7) *Range based (range):* This context defines specific ranges for feature values, using narrow intervals like [x-value, x+value] for continuous features. The size of these intervals, typically small values like 3 or 5, may be informed by domain expertise or other criteria [37]. Categorical/binary features are uniformly distributed. This approach also treats each feature independently and analyses the model's behaviour among samples

with feature values evenly spread within a set range, allowing for an analysis of feature sensitivity.

Using our taxonomy, we can construct complex contexts. For instance, a "prob_range" hybrid context combines the "prob" and "range" contexts. In "prob_range", a sample $x1$ is more likely to be included in the context than $x2$ if $x1$ is nearer to input u, provided both $x1$ and $x2$ fall within a specified feature value range. Samples outside this range are not considered. Merging different neighbourhood contexts allows for a more comprehensive evaluation, leveraging the distinct advantages of each to provide a more detailed analysis.

B. APPLICATION OF CONTEXTS

While multiple contexts can be constructed, not all contexts can be applied to every technique as the methodology to generate scores for features differs. LIME won't function with any restricted contexts (inside/outside) as they require samples from both classes to train their interpretable classifier. SHAP, on the other hand, is more flexible and can be used with restricted contexts as well. Finally, DiCE CFs require restricted contexts outside the decision boundary. Hence, combinations of contexts with an "outside" restriction are used with DiCE.

C. EVALUATION OF EXPLANANDA

This section detail the procedure for each evaluation. The dataset is divided into the training and testing data. The classifier is trained using the training data. Each test sample from the test data x is analysed using XAI frameworks using different context. A modified sample x' with missing features (imputed by their mean values from the training data) represents the scenario with missing information.

For each specified proportion of missing features, i.e., $missing\_proportion$ – 20%, 30%,40%,50%,60% or 70% – 50 trials are conducted to ensure variety in the selection of missing features. The number of missing features is varied as per the percentage of features in datasets because a single feature in diabetes (with less features) can contribute much more than a single feature in cervical cancer (with more features) where as 20% of features in diabetes might be equivalent to the 20% of features in the cervical cancer dataset. As different datasets have different feature types, the complex interaction of features will lead to variation in success rates across datasets. This is discussed further in the results Section V. The code to reproduce results using the proposed methodology is available on Github.[4]

1) EARLY STOPPING

This evaluation examines whether the "present" features are sufficient to maintain the classification, which is a key aspect of FI frameworks. Hence, SHAP and LIME are assessed, while CFs, which do not address sufficiency directly, are not included in this evaluation. For each context and FI framework pair, we calculate FI scores for modified samples (x') and derive feature rankings.

If the average rank of top-k present features is more than that of k missing features, it is interpreted that the critical features have already been gathered, and workflow can be stopped. Successful attempts N_{ES} are recorded by measuring identified cases of early stop that indeed already have a same classification as that of original i.e, N_{ES} is incremented if the classification with mean-imputed missing features $f(x')$ matches the original $f(x)$, implying further data collection is unnecessary. ES_Trials represent the condition of early stopping by counting the total number of trials where the classification with the imputed values in missing features has been the same as the original classification ($f(x') == f(x)$).

Success rates are calculated as the proportion of successful attempts (N_{ES}) to the total number of trials (ES_Trials). This rate indicates the likelihood of successfully predicting early stopping based by comparing FI scores of present and missing features. We evaluate this rate for various contexts to derive an optimal context that will define the sample distribution needed to produce the most useful FI scores for this explanandum. Section V provides this derivation after presenting our results. Statistical significance is determined via a binomial test comparing N_{ES} against ES_Trials, with the hypothesised success probabilities of 50% and a p-value cutoff of 0.005 for significance.

2) NEXT BEST FEATURE

This explanandum evaluates the most important missing feature highlighted by SHAP, LIME and DiCE. CFs from DiCE are included in the evaluation, as they pinpoint features that could alter the classification outcome, specifically toward the accurate classification.

For each combination of context and XAI framework, the next best feature is selected in three distinct ways:

1) *General FI scores (FI'):* The top-ranked missing feature ($TopF$) is chosen from the rankings generated using contexts.
2) *Custom contexts (FI_Cus):* FI scores are recalculated using context c'. In c', the values of the present features are fixed to their original values while varying the 'missing' feature values based on the rules of the specific context. The top-ranked missing feature is then selected ($TopFcus$).
3) *Random selection:* A feature is arbitrarily chosen from the missing set ($RandF$).

When a feature is selected using one of the above-mentioned ways, a successful attempt is noted when restoring the original value of the selected feature - $TopF$, $TopFcus$, and $RandF-$ results in a switch to the original classification (with all values present). The successful attempts using three types of feature selection are noted separately - N_{NBS_Gen}, N_{NBS_Cus}, and N_{NBS_Rand}. The total number of trials where this explanandum is applicable ($f(x') \neq f(x)$) are denoted by NBS_Trials.

[4]Link to the code - https://github.com/UrjaPawar/XAI_Evaluation

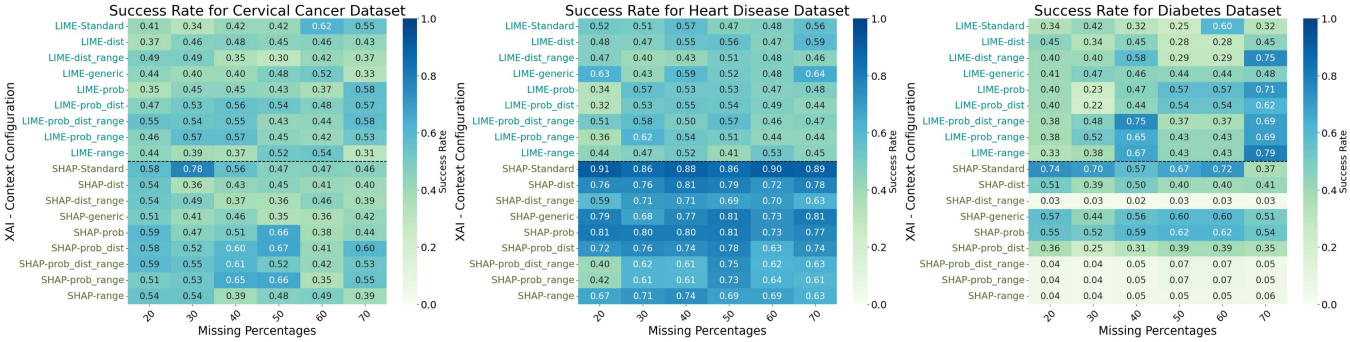

FIGURE 5. Success rates for 'early stopping' using FI frameworks with unrestricted contexts.

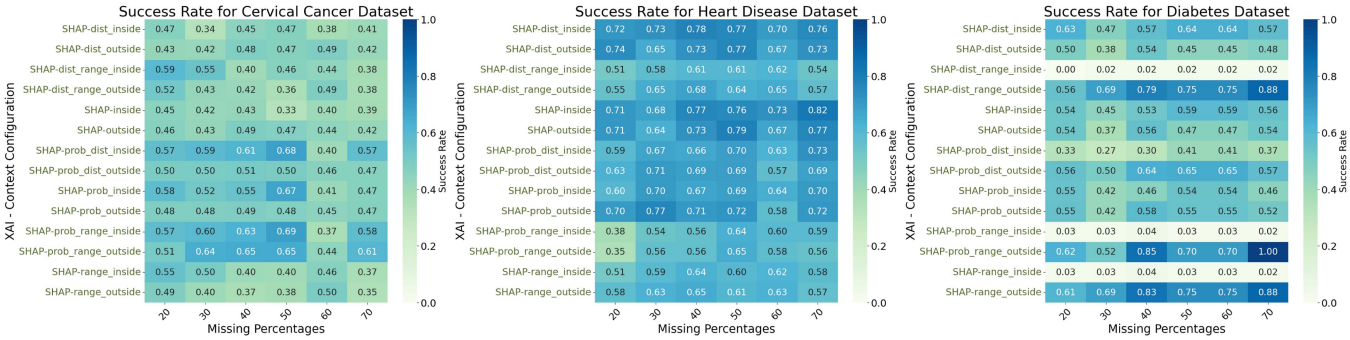

FIGURE 6. Success rates for 'early stopping' using SHAP with restricted contexts.

The success rate in this explanandum is the ratio successful attempts (e.g. N_{NBS}) and NBS_Trials. This rate represents the likelihood of successfully identifying the next best feature based on FI scores of missing features. An optimal context is derived by comparing success rates of various contexts to select samples to generate most useful FI scores. Binomial tests are performed considering the successful attempts and NBS_Trials. The results are considered valid if the success rate is more than 50%. A p-value threshold of 0.005 is used.

V. RESULTS AND DISCUSSIONS

The evaluation assesses how SHAP, LIME, and DiCE perform with different contexts to identify the optimal contexts that suit the selected explananda.

Success rates are visualised using heatmaps: the y-axis displays the XAI framework-context pairs (e.g., SHAP-range), and the x-axis represents percentages of missing features. The colour intensity of each cell indicates the success rate for that particular framework-context pair and dataset, with darker colours denoting higher rates.

A. EARLY STOPPING

This evaluation assesses of FI frameworks in predicting an "early stop". DiCE is not a part of this evaluation.

Fig. 5 shows the success rates of LIME and SHAP on SVM-Linear using various unrestricted contexts, and Fig. 6 shows SHAP's performance in restricted contexts as LIME cannot use these. For the cervical cancer dataset, both SHAP and LIME showed similar effectiveness, but in the heart disease datasets, SHAP outperformed LIME by 50%. In the diabetes dataset, SHAP performed well only in certain contexts. This variation in absolute values of success rates can be attributed to the different numbers and types of missing features in datasets across various missing percentages, indicating that the absolute success rates of XAI frameworks is dependent on the dataset and the missing percentage of features. With increasing percentage of missing features, contexts can generate lesser representative neighborhoods for an input sample with more features imputed. Based on the criterion used for a specific context, the neighborhoods generated will be distinct with each missing percentages that lead to variation in success rates. However, in relative terms, the *standard*, *generic* and contexts based on *prob* in SHAP outperformed others on average across all datasets. The broader sample diversity in these contexts without feature value restrictions enabled capturing the sufficiency of present features. With *range* contexts, SHAP's perturbations failed to create a varied enough sample set for sufficiency analysis.

In LIME, no single context performed well across different missing percentages. However, as the missing percentage increased, on average, contexts based on *range*, such as *prob_range*, generally showed higher success rates across the datasets. With a limited range of values in features, LIME assigns importance as per the model's sensitivity to small changes in feature values. As the number of missing features increased, sensitivity-based feature ranking became more dominant and captured the importance of sensitive missing features. The context *prob* and *prob_dist* also performed well

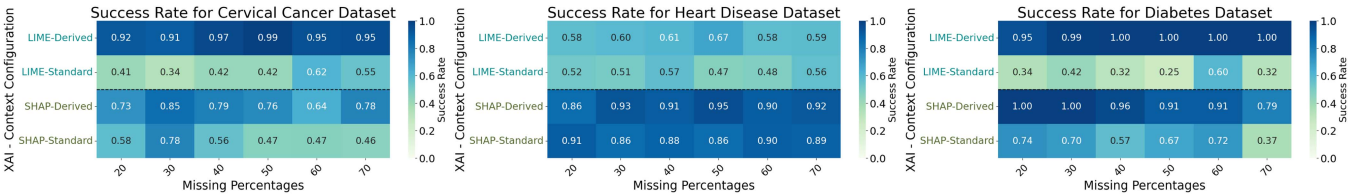

FIGURE 7. Success rates for 'early stopping' using LIME & SHAP with standard and derived contexts.

due to LIME's weighting scheme, which prioritizes samples based on their distance from the original input, and the *prob* context provides more locally similar samples, making importance rankings more reliable. Both *generic* and *standard* contexts provided diverse sets of samples and showed similar success rates across different missing percentages.

The performance of restricted contexts with SHAP is shown in Fig. 6. In the cervical cancer and heart disease dataset, there is no single context that performed best across the missing percentages. In cervical cancer, this is due to a relatively higher number of features in the cervical cancer dataset, it is a difficult to identify a context that addresses both - sufficiency of present features and sensitivity of missing features. In the heart disease dataset, SHAP showed comparable success rates using contexts based on *inside* (averaging 80%) and *outside* (averaging 75%). As there is no classification change in an *inside* context, SHAP assesses feature's importance by analysing difference in original prediction score and the obvious increased score for the current classification that enables sufficiency-based importance ranking. *outside* contexts also performed well as they contain samples of a different class, enabling SHAP to identify sensitive features by measuring how feature alterations can change classification. In the diabetes dataset, as the number of missing features increased to 5 (70% of 7 available features), the *outside* contexts with *range* such as *range_outside*, achieved very high performance with SHAP, averaging 95%. In these contexts, SHAP captures the significance of features based on their sensitivity because all samples belong to the opposite class (*outside*), and have a limited variation in the features' values. As the number of missing features increases, sensitivity-based feature ranking becomes dominant in capturing highly sensitive missing features. The optimal choice between "inside" and "outside" varies by dataset and depends on the specific goal: either to confirm the sufficiency of existing features or to identify sensitive missing features.

As discussed, the sufficiency of present features is captured by assessing the impact of present features' values in strengthening the classification, compared to other values in the data (*standard* or *generic* context). The sensitivity of missing features is best captured in a limited *range*-based context. With this understanding, an optimal context is derived where present features' values are randomly chosen, and missing features are limited to a range. These samples are sorted using the Mahalonobis distance metric, ensuring alignment with training data and the locality. Fig. 7 compares the success rate of this derived context with the standard context of SHAP and LIME. As shown in the figure, the derived context positively affected the performance of both SHAP and LIME, making it useful for this explanandum. The performance gains using the derived contexts over the standard contexts is also validated using RF classifier in Table 3.

TABLE 3. Success Rates for Using RF

Dataset	XAI-context	Missing Percentages 20%	30%	40%	50%	60%	70%
Cervical Cancer	LIME-Derived	**0.91**	**0.90**	**0.93**	**0.97**	**0.95**	**0.94**
	LIME-Standard	0.54	0.28	0.40	0.37	0.51	0.63
	SHAP-Derived	**0.94**	**0.95**	**0.96**	**0.99**	**0.96**	**0.90**
	SHAP-Standard	0.56	0.51	0.43	0.48	0.45	0.40
Diabetes	LIME-Derived	**0.63**	**0.91**	**0.95**	**0.97**	**0.97**	**0.99**
	LIME-Standard	0.40	0.41	0.39	0.33	0.44	0.81
	SHAP-Derived	**0.98**	**0.98**	**0.95**	**0.87**	**0.87**	**0.63**
	SHAP-Standard	0.62	0.57	0.52	0.52	0.54	0.78
Heart Disease	LIME-Derived	0.48	0.50	0.52	**0.62**	0.53	**0.50**
	LIME-Standard	**0.52**	**0.54**	**0.65**	0.48	**0.58**	0.44
	SHAP-Derived	**0.83**	0.87	**0.90**	**0.92**	**0.87**	**0.90**
	SHAP-Standard	0.52	**0.88**	0.80	0.60	0.69	0.83

A binomial test is used to assess the significance of the derived context in SHAP and LIME with the hypothesised success probability of 50%. While *standard* SHAP showed p-values lower than 0.005, the *standard* LIME did not achieve statistically significant success rates, with p-values exceeding 0.005. The p-values for derived contexts in LIME and SHAP were significantly lower than 0.005. The detailed results for RF classifier are provided in the supplementary material (Appendix A). We have further validated the results using median data imputation and the results are included in the supplementary material (Appendix B). In summary, a context defined by a distribution that allows for random values for the present features and only a narrow range of values in imputed missing features is effective in capturing the significance of both sets of features by LIME and SHAP such that their feature ranking can enable the decision of early stopping.

B. NEXT BEST FEATURES

Here, the primary objective is to evaluate how well SHAP, LIME, and DiCE identify the "next best feature" in a workflow, assuming no feature inter-dependency. Success rates are presented using discussed contexts (represented by *general*) and their customised versions with fixed values of present features (represented by *custom*). These results are also compared to rates achieved by random selection.

Figs. 8 and 9 show FI frameworks' performance in *general* and *custom* contexts respectively. Similar to the previous explanandum, the results indicate that the success rate is not

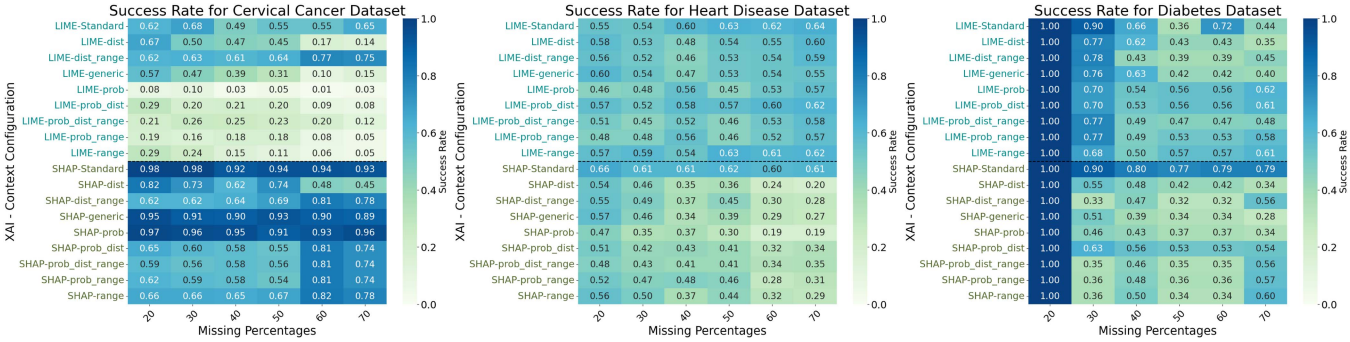

FIGURE 8. Success rates for 'next best feature' using FI frameworks with general unrestricted contexts.

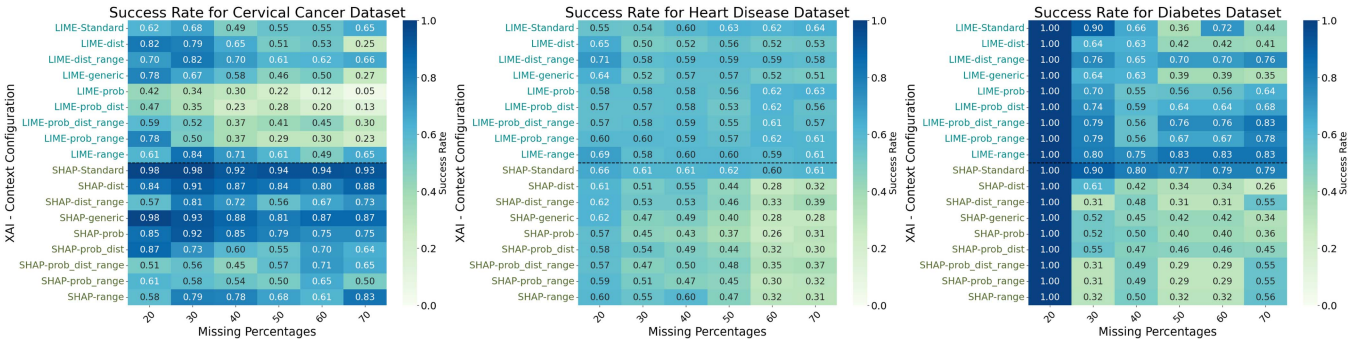

FIGURE 9. Success rates for 'next best feature' using FI frameworks with custom unrestricted contexts.

strongly linked to the percentage of features that are missing. Instead, it depends on the datasets. This is intuitive as the relative rankings among missing features are affected by the total number and type of missing features rather than their proportion to present features.

Fig. 8 displays success rates for LIME and SHAP within *general* contexts. For cervical cancer, SHAP achieved an 80% average success rate, compared to LIME's 25%. In the heart disease dataset, LIME scored better with a 57% average, surpassing SHAP's 40%. In the diabetes dataset, both frameworks had similar performance, averaging around 50%. Notably, success rates are very high when only 20% of features are missing, particularly in the diabetes dataset, where the inclusion of the only one missing feature will result in the accurate classification. This performance variation across datasets can be attributed to various factors such as the number and types of features. However, to derive an optimal context, the relative performances of contexts across the datasets will be analysed. The *standard* context was the only one that consistently performed well by providing samples from the training data that were aligned with the distribution of the test data samples used in the evaluation.

For LIME, there isn't a specific context that proves effective across all datasets. This is because LIME analyses multiple samples with respect to each other, rather than analysing them with respect to the specific input. This is also influenced by the interactions between the missing features' values and the present feature values in other samples. A missing feature that impacts classification in samples with different values of present features, doesn't necessarily indicate the next best feature. Its importance should be evaluated in relation to the current values of the present features. The most insightful context here should enable LIME to analyse classification changes with respect to the missing features.

Fig. 9 shows success rates using customised contexts with fixed values of present features. These contexts significantly improved LIME's performance across datasets. Specifically, there was a 100% improvement in the cervical cancer dataset, 7% in the heart disease dataset, and 20% in the diabetes dataset. Contexts *range* and *dist_range* performed consistently by highlighting the sensitivity of missing features within a range. However, these results have limited statistical significance due to the less number of trials. This is because with limited range of values in missing features and fixed values of present features, there are fewer scenarios where samples from both classes are present for LIME to function. SHAP's performance with custom contexts remained largely the same for the diabetes and heart disease datasets but dropped by 12% in the cervical cancer dataset. This means that apart from *standard* context, SHAP's rankings using *general* as well as *custom* unrestricted contexts are not useful for this explanandum.

Fig. 10 shows the results when *general* restricted contexts are used with SHAP and DiCE. For DiCE, no context performed well across all datasets. In CFs, multiple feature changes make it challenging to identify truly impactful

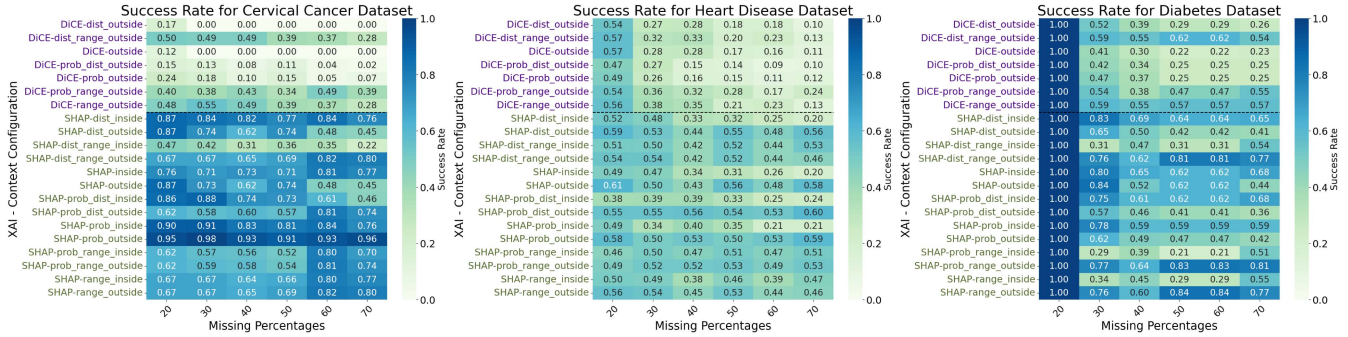

FIGURE 10. Success rates for 'next best feature' using SHAP & DICE with general restricted contexts.

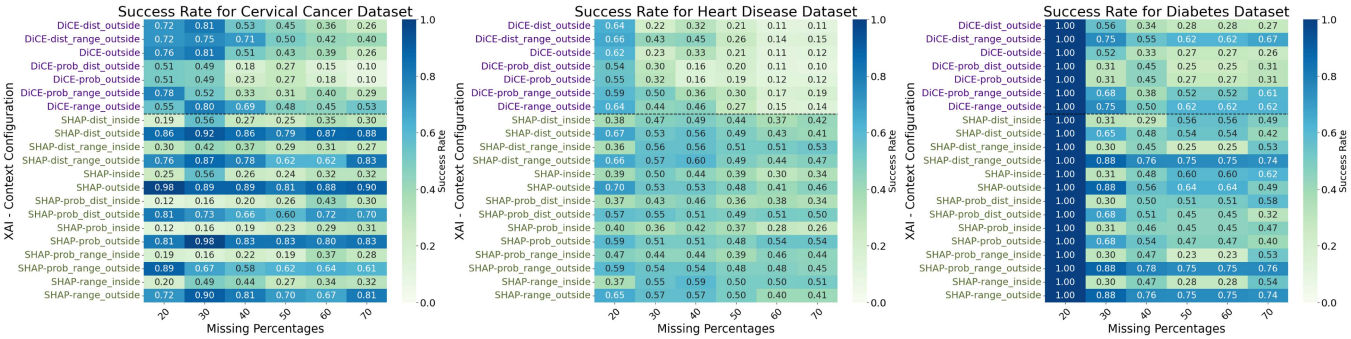

FIGURE 11. Success rates for 'next best feature' using SHAP & DICE with custom restricted contexts.

features. This leads to complicated interactions and makes DiCE CFs ineffective for this explanandum even with *custom* contexts, as shown in Fig. 11.

With *general* restricted contexts(Fig. 10), SHAP's results varied by dataset with no performance patterns related to the *inside* or *outside* contexts. However, when *custom* restricted contexts are used (Fig. 11), a clear demarcation of performance by *outside* is observed in cervical cancer and diabetes datasets. Heart disease showed lower success rates because more than one feature might be necessary to reach the correct classification. This is validated by observing the cumulative performance of top-2 next best features in heart disease using the custom restricted contexts with SHAP in Fig. 12. It indicates that the *outside* context outperformed the *inside* context by 50% on average. The *outside* context contains samples only from the opposite class, making the importance rankings more reflective of key features that could change the classification. This is especially effective with *custom* contexts where fixed present features are used.

The custom *outside* contexts with SHAP did not perform as well as the *standard*, as shown in Fig. 8. For optimal performance, a *custom* context is derived using training data samples by 1) fixing values of *present* features to be same as that in the input sample and 2) selecting only those samples that have a different classification than the input sample. This derived context is hybrid of the *standard* and custom *outside* contexts. The success rates using the derived context in SHAP are shown in Fig. 13; on average, SHAP's performance increased by 12% compared to the standard context. LIME is not used with this context as it is restricted to one class. The

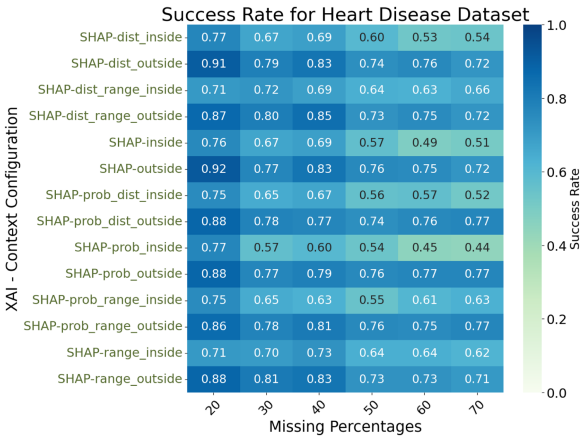

FIGURE 12. Success rates of SHAP using top-2 ranked features with custom restricted contexts.

binomial tests were conducted with the hypothesised success rate of 50%. The statistical significance of the success rates of *derived* contexts with SHAP was high, with p-values significantly lower than 0.005.

To demonstrate the efficacy of FI rankings in this explanandum, the success rates using general and custom contexts were compared to random feature selection as shown in Fig. 14. The three plots correspond to each dataset, with the y-axis showing the average success rate across all the contexts and the x-axis showing the XAI frameworks. The general contexts, custom contexts, and random selection are represented by blue, green, and brown colours, respectively. Darker colours show

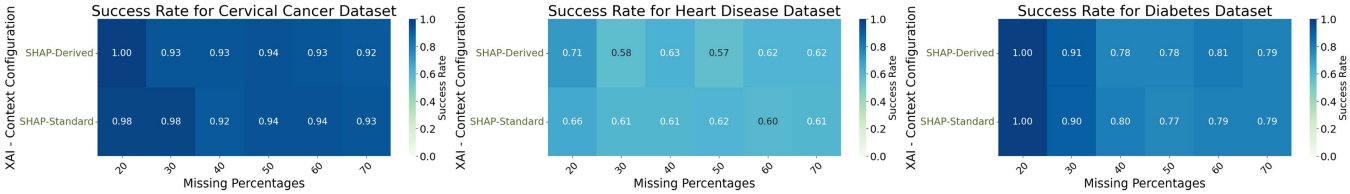

FIGURE 13. Success rates for 'next best feature' using SHAP with the standard and derived contexts.

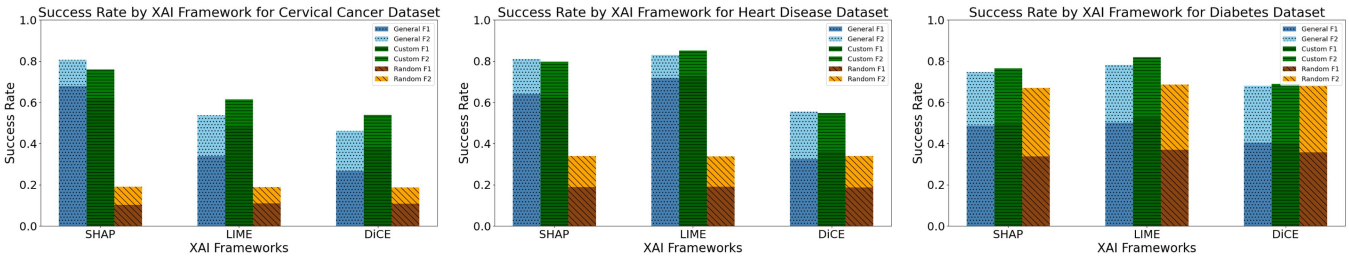

FIGURE 14. Average success rates for 'next best features' using XAI with 1) general, 2) custom, and 3) random contexts.

TABLE 4. Success Rates Using RF

Dataset		Missing Percentages					
	XAI-context	20%	30%	40%	50%	60%	70%
Cervical Cancer	SHAP-Derived	**0.92**	**0.79**	0.79	**0.83**	0.76	**0.77**
	SHAP-Standard	0.79	0.72	**0.80**	0.80	**0.83**	**0.78**
Diabetes	SHAP-Derived	**1.00**	**0.87**	**0.88**	**0.83**	**0.82**	**0.82**
	SHAP-Standard	**1.00**	0.86	0.85	0.81	0.82	0.79
Heart Disease	SHAP-Derived	**0.86**	**0.75**	**0.77**	**0.74**	**0.74**	**0.73**
	SHAP-Standard	0.74	0.75	0.68	0.73	0.70	0.70

success rates when the top-ranked feature is selected, and lighter colours show cumulative gains in success rates when the top-2 features are selected. The plots show that using FI rankings with general/custom contexts are more effective than random selection. In diabetes, random selection also performed well due to the lesser number of features. Overall, FI rankings are more effective than random selection, and using SHAP with the derived context has proven to be particularly effective for this explanandum.

The performance gains using the derived optimal contexts using RF classifier is shown in Table 4. The detailed results are included in supplementary material (Appendix A). The derived context was successfully validated using median data imputation as well and the results are included in Appendix B. In summary, while *custom* contexts improved LIME's performance, the success rates were not always more than 50%. DiCE didn't perform well with any context due to feature interactions. For SHAP, a custom context from training data that is restricted to the samples of different class than the input is effective in generating feature ranking that enables the identification of the next best feature.

VI. RELATED WORK

This section provides an overview of the related work around the significance of explananda and the impact of neighbourhood contexts on the explanations.

Doshi-Velez & Kim [3] categorised XAI evaluations into three types - application-grounded, human-grounded, and functionally-grounded. Our study is an application-grounded evaluation where the focus is to assess the practical impact of explanations in specific use cases of real-world settings. We assess whether provided explanations enable early stopping and identify next best feature. These evaluations involve end-users but can be done without them if the explananda can be simulated and analysed statistically as indicated in [2]. The human-grounded approach evaluates user comprehension and trust in AI through user studies. Within this approach, various frameworks were discussed to tailor the explanations based on the requirements of the audience, such as by using well-defined user studies as proposed in [6], or by evaluating properties such as actionability, and complexity as proposed in [7].

The functionally-grounded evaluation involves a more technical analysis of AI interpretability through metrics such as assessing the faithfulness of explanations to the black-box model, and their granularity (local/global) [3]. For example, ablation studies have been utilised in [8] to observe variation in the model's performance based on the perturbation of features highlighted as "important" by XAI. It was concluded that a decline in performance does not necessarily indicate that the perturbed features are indeed important, as this depends on the perturbation technique and the dataset characteristics. In [9], perturbations were intentionally designed to "trigger" the model into producing incorrect classifications to examine if the influential features according to XAI, were the correct "triggers". While our work is also related to ablation studies, it distinctively focuses on application-grounded evaluation examining neighbourhood contexts to enhance the utility of XAI frameworks.

Designing application-grounded evaluations with a specific explanandum is essential to avoid user misunderstandings. Chromik et al. [4] illustrated this with a study where

participants were asked to self-rate their understanding of an AI system after receiving SHAP explanations and then testing them on mimicking the AI for a set of inputs (the explanandum). The self-ratings decreased significantly, suggesting an illusion of understanding created by the explanations. Similarly, Wang et al. revealed that people's interpretations of explanations to understand the ML model's uncertainty (the explanandum) vary based on their domain expertise [5]. Both studies, however, did not explore how underlying neighbourhood contexts affect the utility of XAI for their explanandum.

A study of explanatory requirements in the medical domain [38] found that clinicians expect explanations to provide them with parsimonious and actionable steps to work efficiently with CDSS. In [14], it is emphasised that CDSS should mimic the clinical decision-making process based on optimising workflows, especially in low-resource settings. While theoretical methodologies for evaluating XAI techniques are discussed in [39], [40], they have not demonstrated potential applicability using experimental evaluations.

Monteath et al. [41] proposed a decision tree-based approach to generate confidence scores while classifying medical records based on available information. These confidence scores were calculated using the incremental information gains achieved as more medical data was added. Notably, the work addressed the two explananda: measuring confidence in the current diagnosis and recommending the next best diagnostic test for individual diagnosis. However, its applicability remained limited to decision tree-based models.

Neighbourhood context plays a critical role in generating local explanations [12], [16]. Ribeiro et al.'s LIME framework demonstrates the impact of modifying samples near the input to better approximate a model's local decision-making [10]. XAI frameworks can also be "fooled" into producing certain explanations through neighbourhood manipulation, illustrating how context selection influences the quality and utility of explanations [11].

VII. LIMITATIONS AND FUTURE WORK

This work introduced a methodology to emphasise the need for explanandum-based evaluations and the impact of neighbourhood contexts. However, it has some limitations. Firstly, there was an assumption of using mean (and median) values to account for missing features instead of considering all possible values in missing features. This also directly impacted the context to be based on imputed values. As we wanted to demonstrate the impact of specific contexts on the utility of FI scores, generating local FI scores for the input samples required the use of imputation for direct applicability at a given stage in a workflow. There was also a lack of specific conditions related to medical workflows for early stopping and the next best feature. This would require an extensive review of medical literature, which is beyond the scope of this work. The method for comparing feature rankings could also be refined. Further, the study didn't explore more complex neighbourhood contexts.

In future work, we will develop methodologies to account for all values in missing features or use specific baseline values as per the medical literature. We will conduct further validation with respect to specific medical workflows in the two explananda for the evaluation. The evaluation will be extended to higher-dimensional datasets and consider more complex variables to draft an explanandum. It will investigate alternative methodologies for comparing feature rankings and use feature-specific scoring mechanisms for generating FI scores using CFs. A broader range of neighbourhood contexts will be explored to refine the effectiveness of XAI for an explanandum. We will also explore optimisation techniques that can be used with XAI frameworks for generating explanations faster [42].

VIII. CONCLUSION

This study presents a methodology to evaluate XAI by formalising two healthcare domain-specific tasks as explananda: 'early stopping' and 'next best feature'. We conducted a rigorous evaluation with various numbers and selections of missing features representing scenarios of each explanandum across different datasets and classifiers. We employed various contexts with popular XAI frameworks - SHAP, LIME, and DiCE, to investigate their utility in addressing the explananda and highlight the impact of neighbourhood contexts on XAI's success rates in each explanandum.

The results obtained revealed interesting insights. Notably, the standard SHAP outperformed the standard versions of LIME and DiCE for both explananda. However, SHAP's performance was improved using more optimal derived contexts. In the 'early stopping', the sufficiency of the values in the present set of features was required to be compared with the sensitivity of the imputed values in the missing features. An optimal context was derived by allowing broad changes in the values of present features but only a narrow range of values in the missing set of features. For the 'next best feature', the classification change due to the missing features was required to be captured. SHAP demonstrated good performance with *outside* contexts, particularly with fixed values of present features, enabling a focused analysis of missing features. An optimal context was derived that combined *standard* and *outside* and showed improved success rates. Statistical tests were done to confirm the significance of these improvements.

Overall, this work facilitates the evaluation of domain-specific utility of XAI with respect to the neighborhood contexts. Our proposed methodology is adaptable to evaluate XAI via proxy application-based explananda by 1) simulating scenarios using different perturbations/imputations and then 2) examining the feature ranking to provide the utility. By highlighting these contributions, we emphasise the practical relevance and broader applicability of our research, offering a promising avenue for advancements in XAI.

REFERENCES

[1] L. Hancox-Li, "Robustness in machine learning explanations: Does it matter?," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 640–647, doi: 10.1145/3351095.3372836.

[2] T. Freiesleben and G. König, "Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research," in *Proc. World Conf. Explainable Artif. Intell.*, 2023, pp. 48–65.

[3] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *stat*, vol. 1050, pp. 1–13, 2017.

[4] M. Chromik, M. Eiband, F. Buchner, and A. Krger, "Butz A. I. Think I. Get your point, AI! the illusion of explanatory depth in explainable AI," in *Proc. 26th Int. Conf. Intell. User Interfaces*, 2021, pp. 307–317, doi: 10.1145/3397481.3450644.

[5] X. Wang and M. Yin, "Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making," in *Proc. 26th Int. Conf. Intell. User Interfaces*, 2021, pp. 318–328, doi: 10.1145/3397481.3450650.

[6] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv.1812.04608*.

[7] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 56–67, doi: 10.1145/3351095.3372870.

[8] I. Hameed et al., "BASED-XAI: Breaking ablation studies down for explainable artificial intelligence," 2022, *arXiv.2207.05566*.

[9] Y. S. Lin, W. C. Lee, and Z. B. Celik, "What do you see? Evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 1027–1035.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[11] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2020, pp. 180–186, doi: 10.1145/3375627.3375830.

[12] T. Han, S. Srinivas, and H. Lakkaraju, "Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 5256–5268.

[13] A. Simkute, E. Luger, B. Jones, M. Evans, and R. Jones, "Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable," *J. Responsible Technol.*, vol. 7, 2021, Art. no. 100017, doi: 10.1016/j.jrt.2021.100017.

[14] D. A. Zikos, "Framework to design successful clinical decision support systems," in *Proc. 10th Int. Conf. Pervasive Technol. Related Assistive Environ.*, 2017, pp. 185–188, doi: 10.1145/3056540.3064960.

[15] D. Kiyasseh, T. Zhu, and D. Clifton, "The promise of clinical decision support systems targetting low-resource settings," *IEEE Rev. Biomed. Eng.*, vol. 15, pp. 354–371, 2022.

[16] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, "Towards unifying feature attribution and counterfactual explanations: Different means to the same end," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2021, pp. 652–663, doi: 10.1145/3461702.3462597.

[17] A. M. Antoniadi et al., "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, 2021, Art. no. 5088, doi: 10.3390/app11115088.

[18] D. S. Watson, L. Gultchin, A. Taly, and L. Floridi, "Local explanations via necessity and sufficiency: Unifying theory and practice," in *Proc. Uncertainty Artif. Intell.*, 2021, pp. 1382–1392.

[19] S. Hooker, D. Erhan, P. jan Kindermans, and B. A. Kim, "Benchmark for interpretability methods in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 9737–9748.

[20] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv JL Tech.*, vol. 31, 2017, Art. no. 841.

[21] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[22] S. Mishra, S. Dutta, J. Long, and D. Magazzeni, "A survey on the robustness of feature importance and counterfactual explanations," 2021, doi: 10.48550/arXiv.2111.00358.

[23] I. Čík, A. D. Rasamoelina, M. Mach, and P. Sinčák, "Explaining deep neural network using layer-wise relevance propagation and integrated gradients," in *Proc. IEEE 19th World Symp. Appl. Mach. Intell. Inform.*, 2021, pp. 000381–000386, doi: 10.1109/SAMI50585.2021.9378686.

[24] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2017, vol. 30, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[25] S. Palacio, A. Lucieri, M. Munir, S. Ahmed, J. Hees, and A. Dengel, "XAI handbook: Towards a unified framework for explainable AI," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3766–3775.

[26] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 607–617, doi: 10.1145/3351095.3372850.

[27] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Syst. With Appl.*, vol. 42, no. 13, pp. 5621–5631, 2015.

[28] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, pp. 1–8, 2016.

[29] A. Kalyakulina, I. Yusipov, M. G. Bacalini, C. Franceschi, M. Vedunova, and M. Ivanchenko, "Disease classification for whole-blood DNA methylation: Meta-analysis, missing values imputation, and XAI," *Gigascience*, vol. 11, 2022, Art. no. giac097.

[30] M. Thelwall, "The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach," *J. Inform.*, vol. 10, no. 1, pp. 110–123, 2016, doi: 10.1016/j.joi.2015.12.001.

[31] U. Pawar, C. T. Culbert, and R. O'Reilly, "Evaluating hierarchical medical workflows using feature importance," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst.*, 2021, pp. 265–270.

[32] J. S. C. Kelwin Fernandes and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, Springer International Publishing, 2017, pp. 243–250. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cervicalcancer

[33] Robert Detrano PD M D. Uci. 2010. V. A. Medical Center, Long Beach and Cleveland Clinic Foundation, 2010 [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heartdisease

[34] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Symp. Comput. Appl. Med. Care*, IEEE Computer Society Press, 1988, pp. 261–265. [Online]. Available: https://data.world/uci/pima-indians-diabetes

[35] Y. L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Inf. Fusion*, vol. 81, pp. 59–83, 2022.

[36] A. V. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, Springer, 2021, pp. 650–665, doi: 10.1007/978-3-030-86520-7_40.

[37] N. Wang et al., "Study on the semi-supervised learning-based patient similarity from heterogeneous electronic medical records," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. Suppl 2, pp. 1–13, 2021, doi: 10.1186/s12911-021-01432-x.

[38] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proc. Mach. Learn. Healthcare Conf.*, 2019, pp. 359–380.

[39] T. Vermeire, T. Laugel, X. Renard, D. Martens, and M. Detyniecki, "How to choose an explainability method? Towards a methodical implementation of XAI in practice," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2021, pp. 521–533.

[40] A. Adhikari, E. Wenink, J. van der Waa, C. Bouter, I. Tolios, and S. Raaijmakers, "Towards FAIR explainable AI: A standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems," in *Proc. 15th Int. Conf. Pervasive Technol. Related Assistive Environ.*, 2022, pp. 562–568, doi: 10.1145/3529190.3535693.

[41] I. Monteath and R. Sheh, "Assisted and incremental medical diagnosis using explainable artificial intelligence," in *Proc. 2nd Workshop Explainable Artif. Intell.*, 2018, pp. 104–108.

[42] N. Jethani, M. Sudarshan, I. C. Covert, S. I. Lee, and R. Ranganath, "Fastshap: Real-time shapley value estimation," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–12.

URJA PAWAR received the B.Tech. degree in information technology from the National Institute of Technology, Raipur, India. She is currently working toward the Ph.D. degree with Munster Technological University, Cork, Ireland, focusing on Explainable AI. She was a Software Engineer with Optum Global Solutions. She has authored or coauthored several research papers on applications of AI in the medical domain, and explainability of AI/ML models.

DONNA O'SHEA (Senior Member, IEEE) is currently the Chair of Cybersecurity with Munster Technological University, Cork, Ireland. She is a co-Principal Investigator with the Science Foundation Ireland Research Centres CONFIRM and CONNECT, and leads the Ríomh–Intelligent Secure Systems Research Group. She also directs a national project addressing cybersecurity skills challenges, is actively involved in promoting technology careers for all, and the Director of IT at Cork.

RUAIRI O'REILLY (Senior Member, IEEE) received the Ph.D. degree from University College Cork, Cork, Ireland, with research focused on monitoring time series data. He is currently a distinguished computer science academic with Munster Technological University (MTU), Cork. He specialising in artificial intelligence, pattern recognition, and distributed systems with MTU. He teaches knowledge representation and reasoning, and interactive data visualization, while his research enhances e-health workflows and explores affective computing in applied psychology. He supervises Ph.D. students in explainable AI and biomedical image analysis, actively contributes to international conferences, and secures funding from prominent agencies like Science Foundation Ireland and Horizon 2020.

MAEBH COSTELLO is currently the Senior Director of Digital Experience and Design with McKesson, Cork, Ireland, where she has been instrumental since 2020 in leading digital transformations at their County Cork, Ireland office. She has driven significant organisational and cultural shifts by implementing a comprehensive digital strategy that enhances customer engagement through intuitive and seamless interfaces across all platforms. Her responsibilities include overseeing product management, pioneering a data strategy that leverages ML/AI for optimized operations, and fostering collaboration across departments to align digital initiatives with broader corporate goals.

CHRISTIAN BEDER is currently a Lecturer with Computer Science Department, Munster Technological University, where he is also the course coordinator for B.Sc. (Hons.) in software development. He teaches a range of subjects including data structures and algorithms, machine learning, computer vision, and combinatorial optimization across both B.Sc. (Hons.) software development and M.Sc. in artificial intelligence programs. His research interests include several advanced areas: computer vision and photogrammetry, model-based optimization, explainable AI, machine learning, statistical pattern recognition, and their applications in human-cyber-physical systems.